International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-5 Issue-5, June 2016

A Survey on Big Data Analysis and Challenges

J. Samatha, K. Bhagya Laxmi

Abstract- One of contemporary big challenges in information systems is the issues associated with coping with and utilization of vast amounts of data. In this paper we present applications of big data , analysis of big data. The analysis of big data involves phases such as acquisition / recording, extraction / cleaning / annotation, integration / aggregation / representation, analysis / modeling, interpretation. We also discuss the challenges introduced in these phases.

Keywords- Bigdata, volume, velocity, variety, extraction, integration, analysis.

I. INTRODUCTION

Today we live in the digital world. With increased digitization [1] the amount of structured and unstructured data being created and stored is exploding. The data is being generated from various sources –transactions, social media, sensors, digital images, videos, audios and click streams for domains including health care, retail, energy and utilities. In addition to business and organizations [3], individuals contribute to the data volume. For instance, 30 billion content are being shared on Face book every month, the photos viewed every 16 seconds in Picasa could cover a football field.

It gets more interesting. IDC terms this as the 'Digital Universe' and predicts that this digital universe is set to explode to an unimaginable 8 Zeta bytes by the year 2015. This would roughly be a stack of DVD's Earth all the way to Mars. The term "Big Data" was coined to address this massive volume of data storage and processing [6].

A. What Is Big Data

Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools. It is increasingly becoming imperative for organizations[1] to mine this data to stay competitive. Analyzing data can provide significant competitive advantage for an enterprise. The data when analyzed properly leads to a wealth of information which helps the businesses to redefine strategies. However the current volume of big data sets are too complicated to be managed and processed by conventional relational databases and data warehousing technologies.

Manuscript published on 30 June 2016.
***** Correspondence Author (s)
J. Samatha, Department of Computer Science & Engineering, Matrusri Engineering College, Hyderabad (Telangana). India.
K.Bhagya Laxmi, Department of Computer Science & Engineering, Matrusri Engineering College, Hyderabad (Telangana). India.

II. CHARACTERIZATION OF BIG DATA

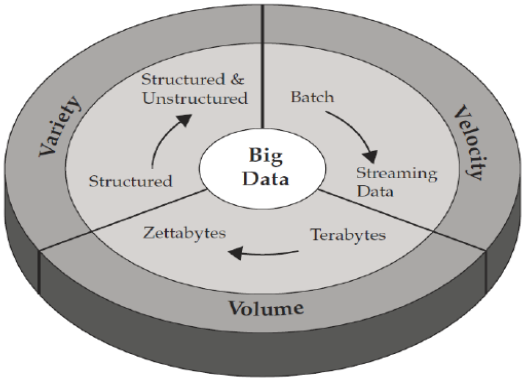

Fig 1: Characterization of big data

The three Vs [5]of volume, velocity and variety are commonly used to characterize different aspects of big data as in Fig1. They're a helpful lens through which to view and understand the nature of the data and the software platforms available to exploit them.

These Vs of big data causes performance problems when being created managed and analyzed using the conventional data processing techniques. Using conventional techniques for big data storage and analysis is less efficient as memory access is slower. The data collection is also challenging as the volume and variety of data has to be derived from sources of different types. The other major challenge in using the existing techniques is they require high end hardware to handle the data with a huge volume, velocity and variety.

A. Velocity

The Velocity is the speed at which the data is created, stored, analyzed and visualized. In the past, when batch processing was common practice, it was normal to receive an update from the database every night or even every week. Computers [3] and servers required substantial time to process the data and update the databases. In the big data era, data is created in real-time or near real-time. With the availability of Internet connected devices, wireless or wired, machines and devices can pass-on their data the moment it is created.

The speed at which data [3] is created currently is almost unimaginable: Every minute we upload 100 hours of video on YouTube. In addition, every minute over 200 million emails are sent, around 20 million photos are viewed and 30.000 uploaded on Flicker, almost 300.000 tweets are sent and almost 2,5 million queries on Google are performed. The challenge organizations have is to cope with the enormous speed the data is created and used in real-time.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

B. Volume

90% of all data ever created, was created in the past 2 years. From now on, the amounts of data in the world [4] will double every two years. By 2020, we will have 50 times the amount of data as that we had in 2011. The sheer volume of the data is enormous and a very large contributor to the ever expanding digital universe is the Internet of Things with sensors all over the world in all devices creating data every second. The era of a trillion sensors is upon us.

If we look at airplanes they generate approximately 2.5 billion Terabyte of data each year from the sensors installed in the engines. Self-driving cars will generate 2 peta bytes of data every year. Also the agricultural industry generates massive amounts of data with sensors installed in tractors. Shell uses super-sensitive sensors to find additional oil in wells and if they install these sensors at all 10.000 wells they will collect approximately 10 Exabyte of data annually. That again is absolutely nothing if we compare it to the Square Kilometer Array Telescope that will generate Exabyte of data per day.

In the past, the creation of so much data would have caused serious problems. Nowadays, with decreasing storage costs, better storage solutions like Hadoop and the algorithms to create meaning from all that data this is not a problem at all.

C. Variety

In the past[1], all data that was created was structured data, it neatly fitted in columns and rows but those days are over. Nowadays, 90% of the data that is generated by organization is unstructured data. Data today comes in many different formats: structured data, semi-structured data, unstructured data and even complex structured data. The wide variety of data requires a different approach as well as different techniques to store all raw data.

There are many different types of data and each of those types of data requires different types of analyses or different tools to use. Social media like face book posts or Tweets can give different insights, such as sentiment analysis on your brand, while sensory data will give you information about how a product is used and what the mistakes are.

III. APPLICATIONS OF BIG DATA

Every aspect of our lives will be affected by big data. However, there are some areas where big data is already making a real difference today.

A. Understanding and Targeting Customers

This is [7] one of the biggest and most publicized areas of big data use today. Here, big data is used to better understand customers and their behaviors and preferences. Companies are keen to expand their traditional data sets with social media data, browser logs as well as text analytics and sensor data to get a more complete picture of their customers. The big objective, in many cases, is to create predictive models. Using big data, Telecom companies can now better predict customer churn. Car insurance companies understand how well their customers actually drive. Even government election campaigns can be optimized using big data analytics.

B. Understanding and Optimizing Business Processes

Big data is also increasingly used to optimize business processes. Retailers are able to optimize their stock based on predictions generated from social media data, web search trends and weather forecasts. One particular business process that is seeing a lot of big data analytics is supply chain or delivery route optimization. Here, geographic positioning and radio frequency identification sensors are used to track goods or delivery vehicles and optimize routes by integrating live traffic data, etc. HR business processes are also being improved using big data analytics.

C. Personal Quantification and Performance Optimization

Big data is not just for companies and governments but also for all of us individually[7]. We can now benefit from the data generated from wearable devices such as smart watches or smart bracelets. The other area where we benefit from big data analytics is finding love - online this is. Most online dating sites apply big data tools and algorithms to find us the most appropriate matches.

D. Improving Healthcare and Public Health

The computing power of big data analytics enables us to decode entire DNA strings in minutes[8] and will allow us to find new cures and better understand and predict disease patterns. By recording and analyzing every heart beat and breathing pattern of every baby, the unit was able to develop algorithms that can now predict infections 24 hours before any physical symptoms appear. That way, the team can intervene early and save fragile babies in an environment where every hour counts. What's more, big data analytics allow us to monitor and predict the developments of epidemics and disease outbreaks. Integrating data from medical records with social media analytics enables us to monitor flu outbreaks in real-time.

E. Improving Science and Research

Science and research is currently being transformed by the new possibilities big data brings. CERN, the Swiss nuclear physics lab with its Large Hadron Collider, the world's largest and most powerful particle accelerator. Experiments to unlock the secrets of our universe – how it started and works [1]- generate huge amounts of data. The CERN data center has 65,000 processors to analyze its 30 petabytes of data. However, it uses the computing powers of thousands of computers distributed across 150 data centers worldwide to analyze the data. Such computing powers can be leveraged to transform so many other areas of science and research.

F. Optimizing Machine and Device Performance

Big data analytics help machines and devices become smarter and more autonomous. Big data tools are used to operate Google's self-driving car. The Toyota Prius is fitted with cameras, GPS as well as powerful computers and sensors to safely drive on the road without the intervention of human beings.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Big data tools are also used to optimize energy grids using data from smart meters. We can even use big data tools to optimize the performance of computers and data warehouses.

G. Improving Security and Law Enforcement

Big data is applied heavily in improving security and enabling law enforcement. Security agencies use big data techniques to detect and prevent cyber attacks. Police forces use big data tools to catch criminals and even predict criminal activity and credit card companies use big data use it to detect fraudulent transactions.

H. Improving and Optimizing Cities and Countries

Big data is used to improve many aspects of our cities and countries. It allows cities to optimize traffic flows based on real time traffic information[6] as well as social media and weather data. A number of cities are currently piloting big data analytics with the aim of turning themselves into Smart Cities, where the transport infrastructure and utility processes are all joined up. Where a bus would wait for a delayed train and where traffic signals predict traffic volumes and operate to minimize jams.

I. Financial Trading

High-Frequency Trading (HFT) is an area where big data finds a lot of use today. Here, big data algorithms are used to make trading decisions. Today, the majority of equity trading now takes place via data algorithms that increasingly take into account signals from social media networks and news websites to make buy and sell decisions in split seconds.

IV. ANALYSIS OF BIG DATA

The analysis of Big Data involves multiple distinct phases as shown in the fig2, each of which introduces challenges.

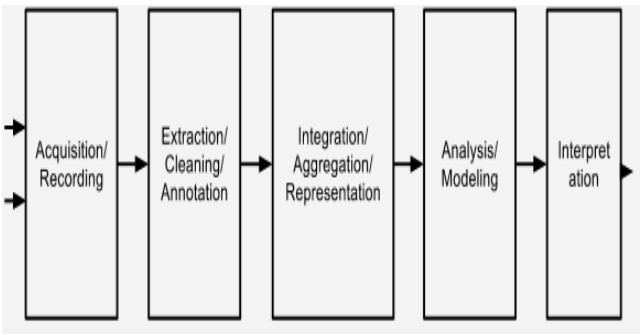

Fig 2: Big data analysis pipeline

A. Data Acquisition and Recording

Big Data does not arise out of a vacuum: it is recorded from some data generating source. For example, consider our ability to sense and observe the world around us, from the heart rate of an elderly citizen, and presence of toxins in the air we breathe, to the planned square kilometer array telescope, which will produce up to 1 million terabytes of raw data per day. Similarly, scientific experiments and simulations can easily produce petabytes of data today. Much of this data is of no interest, and it can be filtered and compressed by orders of magnitude. The filters are defined in such a way that they do not discard useful information.

B. Information Extraction and Cleaning

Frequently, the information collected will not be in a format ready for analysis. For example, consider the collection of electronic health records in a hospital, comprising transcribed dictations from several physicians, structured data from sensors and measurements, and image data such as x-rays[3]. We cannot leave the data in this form and still effectively analyze it. Rather we require an information extraction process that pulls out the required information from the underlying sources and expresses it in a structured form suitable for analysis.

C. Data Integration, Aggregation, and Representation

Given the heterogeneity of the flood of data, it is not[4] enough merely to record it and throw it into a repository. Consider, for example, data from a range of scientific experiments. If we just have a bunch of data sets in a repository, it is unlikely anyone will ever be able to find, let alone reuse, any of this data. Data analysis is considerably more important than simply locating, identifying, understanding, and citing data. For effective large-scale analysis all of this has to happen in a completely[5] automated manner. This requires differences in data structure and semantics to be expressed in forms that are computer understandable, and then "robotically" resolvable. There is a strong body of work in data integration that can provide some of the answers. However, considerable additional work is required to achieve automated error-free difference resolution. Even for simpler analyses that depend on only one data set, there remains an important question of suitable database design. Usually, there will be many alternative ways in which to store the same information. Certain designs will have advantages over others for certain purposes, and possibly drawbacks for other purposes.

D. Analysis / Modeling

Query Processing, Data Modeling, and [6]Analysis Methods for querying and mining Big Data are fundamentally different from traditional statistical analysis on small samples. Big Data is often noisy, dynamic, heterogeneous, inter-related and untrustworthy. Nevertheless, even noisy Big Data could be more valuable than tiny samples because general statistics obtained from frequent patterns and correlation analysis usually overpower individual fluctuations and often disclose more reliable hidden patterns and knowledge. Further, interconnected Big Data forms large heterogeneous information networks, with which information redundancy can be explored to compensate for missing data, to crosscheck conflicting cases, to validate trustworthy relationships, to disclose inherent clusters, and to uncover hidden relationships and models. Mining requires integrated, cleaned, trustworthy, and efficiently accessible data, declarative query and mining interfaces, scalable mining algorithms, and big-data computing environments.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

At the same time, data mining itself can also be used to help improve the quality and trustworthiness of the data, understand its semantics, and provide intelligent querying functions.

E. Interpretation

Having the ability to analyze[7] Big Data is of limited value if users cannot understand the analysis. Ultimately, a decision-maker, provided with the result of analysis, has to interpret these results. This interpretation cannot happen in a vacuum. Usually, it involves examining all the assumptions made and retracing the analysis. Furthermore, there are many possible sources of error: computer systems can have bugs, models almost always have assumptions, and results can be based on erroneous data. For all of these reasons, no responsible user will cede authority to the computer system.

In short, it is rarely enough to provide just the results. Rather, one must provide supplementary information that explains how each result was derived, and based upon precisely what inputs. Such supplementary information is called the provenance of the (result) data.

V. CHALLENGES IN BIG DATA ANALYSIS

Having described[2] the multiple phases in the Big Data analysis pipeline, we now turn to some common challenges that underlie many, and sometimes all, of these phases.

A. Heterogeneity and Incompleteness

When humans consume information, a great deal of heterogeneity is comfortably tolerated. However, machine analysis algorithms expect homogeneous data. In consequence, data must be carefully structured as a first step in data analysis. Consider, for example, a patient who has multiple medical procedures at a hospital. We could create one record per medical procedure or laboratory test, one record for the entire hospital stay, or one record for all lifetime hospital interactions of this patient. With anything other than the first design, the number of medical procedures and lab tests per record would be different for each patient. The three design choices listed have successively less structure and, conversely, successively greater variety[2]. Greater structure is likely to be required by many (traditional) data analysis systems. However, the less structured design is likely to be more effective for many purposes – for example questions relating to disease progression over time will require an expensive join operation with the first two designs, but can be avoided with the latter. However, computer systems work most efficiently if they can store multiple items that are all identical in size and structure. Efficient representation, access, and analysis of semi-structured data require further work.

Consider an electronic health record database design that has fields for birth date, occupation, and blood type for each patient. What do we do if one or more of these pieces of information is not provided by a patient? Obviously, the health record is still placed in the database, but with the corresponding attribute values being set to NULL. [1]A data analysis that looks to classify patients by, say, occupation, must take into account patients for which this information is not known. Worse, these patients with unknown occupations can be ignored in the analysis only if we have reason to believe that they are otherwise statistically similar to the patients with known occupation for the analysis performed. For example, if unemployed patients are more likely to hide their employment status, analysis results may be skewed in that it considers a more employed population mix than exists, and hence potentially one that has differences in occupation-related health-profiles.

Even after data cleaning and error correction, some incompleteness and some errors in data are likely to remain. This incompleteness and these errors must be managed during data analysis. Doing this correctly is a challenge.

B. Scale

Of course, the first thing anyone thinks of with Big Data is its size. After all, the word “big” is there in the very name. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by processors getting faster, following Moore’s law, to provide us with the resources needed to cope with increasing volumes of data. But, there is a fundamental shift underway now: data volume is scaling faster than compute resources, and CPU speeds are static.

First, over the last five years[8] the processor technology has made a dramatic shift - rather than processors doubling their clock cycle frequency every 18-24 months, now, due to power constraints, clock speeds have largely stalled and processors are being built with increasing numbers of cores. In the past, large data processing systems had to worry about parallelism across nodes in a cluster; now, one has to deal with parallelism within a single node. Unfortunately, parallel data processing techniques that were applied in the past for processing data across nodes don’t directly apply for intra-node parallelism, since the architecture looks very different; for example, there are many more hardware resources such as processor caches and processor memory channels that are shared across cores in a single node. Furthermore, the move towards packing multiple sockets (each with 10s of cores) adds another level of complexity for intra-node parallelism. Finally, with predictions of “dark silicon”, namely that power consideration will likely in the future prohibit us from using all of the hardware in the system continuously, data processing systems will likely have to actively manage the power consumption of the processor. These unprecedented changes require us to rethink how we design, build and operate data processing components.

The second dramatic shift that is underway is the move towards cloud computing, which now aggregates multiple disparate workloads with varying performance goals into very large clusters. This level of sharing of resources on expensive and large clusters requires new ways of determining how to run and execute data processing jobs so that we can meet the goals of each workload cost-effectively, and to deal with system failures, which occur more frequently as we operate on larger and larger clusters (that are required to deal with the rapid growth in data volumes).

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

This places a premium on declarative approaches to expressing programs, even those doing complex machine learning tasks, since global optimization across multiple users' programs is necessary for good overall performance. Reliance on user-driven program optimizations is likely to lead to poor cluster utilization, since users are unaware of other users' programs. System-driven holistic optimization requires programs to be sufficiently transparent, e.g., as in relational database systems, where declarative query languages are designed with this in mind.

A third dramatic shift that is underway is the transformative change of the traditional I/O subsystem. For many decades, hard disk drives (HDDs) were used to store persistent data. HDDs had far slower random IO performance than sequential IO performance, and data processing engines formatted their data and designed their query processing methods to "work around" this limitation. But, HDDs are increasingly being replaced by solid state drives today, and other technologies such as Phase Change Memory are around the corner. These newer storage technologies do not have the same large spread in performance between the sequential and random I/O performance, which requires a rethinking of how we design storage subsystems for data processing systems. Implications of this changing storage subsystem potentially touch every aspect of data processing, including query processing algorithms, query scheduling, database design, concurrency control methods and recovery methods.

C. Timeliness

The larger the data[2] set to be processed, the longer it will take to analyze. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster. However, it is not just this speed that is usually meant when one speaks of Velocity in the context of Big Data.

There are many situations in which the result of the analysis is required immediately. For example, if a fraudulent credit card transaction is suspected, it should ideally be flagged before the transaction is completed – potentially preventing the transaction from taking place at all. Obviously, a full analysis of a user's purchase history is not likely to be feasible in real-time. Rather, we need to develop partial results in advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination.

Given a large data set, it is often necessary to find elements in it that meet a specified criterion. In the course of data analysis, this sort of search is likely to occur repeatedly. Scanning the entire data set to find suitable elements is obviously impractical. Rather, index structures are created in advance to permit finding qualifying elements quickly. The problem is that each index structure is designed to support only some classes of criteria. With new analyses desired using Big Data, there are new types of criteria specified, and a need to devise new index structures to support such criteria. For example, consider a traffic management system with information regarding thousands of vehicles and local hot spots on roadways. [8]The system may need to predict potential congestion points along a route chosen by a user, and suggest alternatives. Doing so requires evaluating multiple spatial proximity queries working with the trajectories of moving objects. New index structures are required to support such queries. Designing such structures becomes particularly challenging when the data volume is growing rapidly and the queries have tight response time limits.

D. Privacy

The privacy of data is another huge concern, and one that increases in the context of Big Data. For electronic health records, there are strict laws governing what can and cannot be done. For other data, regulations, are less forceful. However, there is great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources. Managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of big data.

Consider, for example, data gleaned from location-based services. These new architectures require a user to share his/her location with the service provider, resulting in obvious privacy concerns. Note that hiding the user's identity alone without hiding her location would not properly address these privacy concerns. An attacker or a location-based server can infer the identity of the query source from its location information. One types of private information such as health issues or religious preferences can also be revealed by just observing anonymous users' movement and usage pattern over time. Hiding a user location is much more challenging than hiding his/her identity. This is because with location-based services, the location of the user is needed for a successful data access or data collection, while the identity of the user is not necessary. There are many additional challenging research problems. For example, we do not know yet how to share private data while limiting disclosure and ensuring sufficient data utility in the shared data. The existing paradigm of differential privacy is a very important step in the right direction, but it unfortunately reduces information content too far in order to be useful in most practical cases. In addition, real data is not static but gets larger and changes over time; none of the prevailing techniques results in any useful content being released in this scenario. Yet another very important direction is to rethink security for information sharing in Big Data use cases. Many online services today require us to share private information (think of Facebook applications), but beyond record-level access control we do not understand what it means to share data, how the shared data can be linked, and how to give users fine-grained control over this sharing.

E. Human Collaboration

In spite of the tremendous [2] advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a hard time finding. Indeed, CAPTCHAs exploit precisely this fact to tell human web users apart from computer programs.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Ideally, Analytics for Big Data will not be all computational – rather it will be designed explicitly to have a human in the loop. The new sub-field of visual analytics is attempting to do this, at least with respect to the modeling and analysis phase in the pipeline. There is similar value to human input at all stages of the analysis pipeline.

In today's complex world, it often takes multiple experts from different domains to really understand what is going on. A Big Data analysis system must support input from multiple human experts, and shared exploration of results. These multiple experts may be separated in space and time when it is too expensive to assemble an entire team together in one room. The data system has to accept this distributed expert input, and support their collaboration.

A popular new method of harnessing human ingenuity to solve problems is through crowd-sourcing. Wikipedia, the online encyclopedia, is perhaps the best known example of crowd-sourced data. We are relying upon information provided by unvented strangers. Most often, what they say is correct. However, we should expect there to be individuals who have other motives and abilities – some may have a reason to provide false information in an intentional attempt to mislead. While most such errors will be detected and corrected by others in the crowd, we need technologies to facilitate this. We also need a framework to use in analysis of such crowd-sourced data with conflicting statements. As humans, we can look at reviews of a restaurant, some of which are positive and others critical, and come up with a summary assessment based on which we can decide whether to try eating there. We need computers to be able to do the equivalent. The issues of uncertainty and error become even more pronounced in a specific type of crowd-sourcing, termed participatory-sensing. In this case, every person with a mobile phone can act as a multi-modal sensor collecting various types of data instantaneously (e.g., picture, video, audio, location, time, speed, direction, acceleration). The extra challenge here is the inherent uncertainty of the data collection devices. The fact that collected data are probably spatially and temporally correlated can be exploited to better assess their correctness. When crowd-sourced data is obtained for hire, such as with "Mechanical Turks," much of the data created may be with a primary objective of getting it done quickly rather than correctly. This is yet another error model, which must be planned for explicitly when it applies.

VI. CONCLUSION

We have entered an era of Big Data. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. We must support and encourage fundamental research towards addressing these technical challenges.

REFERENCES

1. E.Dumbill, "what is big data? An introduction to the big data landscape", Strata O'Reilly, 11 January 2012.
2. David Loshin, Addressing five emerging challenges of big data, whitepaper.
3. Marko Grobelnik, "Big data tutorial",Stavanger,8 May 2012.
4. Oracle enterprise architecture white paper "An enterprise architect's guide to big data" May 2015.
5. Amir H. Payberah "Introduction to big data", Swedish institute of computer science, 8 April 2014.
6. www.intel.com/bigdata
7. Kostas Glinos, "E-infrastructures for bigdata" ERCIM news, number 89, April 2012.
8. Silva Robak, Bogdan Franczyk, Marcin Robak "Research problems associated with big data utilization in logistics and supply chains design and management" ACSIS, Vol 3,2014

J. Samatha, Asst. Professor, CSE Dept in Matrusri Engineering College, Hyderabad. She is M.Tech(CSE) from JNTU is in teaching field from past ten years. Her interested areas are Cloud Computing, Big Data, Operating Systems.

K. Bhagya Laxmi Asst. Professor, CSE Dept in Matrusri Engineering College, Hyderabad. She is M.Tech (CSE) from JNTU is in teaching field from past ten years. Her interested areas are Cloud Computing, Big Data, Data Mining.